Australian Government
Department of Defence
Defence Science and
Technology Organisation

A Confidence Estimator for Speaker Verification Using Dual DET Curves

T. C. Tao

Command, Control, Communications and Intelligence Division

Defence Science and Technology Organisation

DSTO–RR–0358

ABSTRACT

In speaker verification, the result of a trial is traditionally summarised as an arbitrary score, where a higher score indicates stronger evidence in favour of the speaker hypothesis. However this is difficult to interpret. It is useful to convert this score into a "confidence level", i.e. the posterior probability that the speaker hypothesis is correct, given the score. One of the simplest formulae to obtain a confidence level is using a logistic curve, but this requires the assumption that the true and impostor speaker scores are distributed according to a Normal distribution. In this report I propose a new formula, called the dual Detection Error Trade-Off (DET) curve, since it represents the same information as a DET curve. This formula avoids the assumption of normally distributed target and impostor scores. Experiments on the NIST 99 data prove the dual DET curve performs slightly better than the logistic curve.

APPROVED FOR PUBLIC RELEASE

Published by

DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh, South Australia 5111, Australia

Telephone: (08) 8259 5555
Facsimile: (08) 8259 6567

© *Commonwealth of Australia 2010*
AR No. AR-014-858
October, 2010

APPROVED FOR PUBLIC RELEASE

A Confidence Estimator for Speaker Verification Using Dual DET Curves

Executive Summary

Speaker recognition is the problem of identifying people from their voices, and has important applications. For instance, it is often desirable to determine (or confirm in the case of Speaker *Verification*) the identities of various speakers in various applications, such as telephone calls. The human voice is unique, or at least very difficult to mimic successfully, so the ability to identify a speaker based on his or her voice offers better security over some other means of identification such as passwords (which can be forgotten or compromised) or physical objects (which can be stolen).

In speaker verification, the result of a trial is traditionally summarised as an arbitrary score, where a higher score indicates stronger evidence in favour of the speaker hypothesis. However this is difficult to interpret. It is useful to convert this score into a "confidence level", i.e. the a-posteriori probability that the speaker hypothesis is correct, given the score. Ideally, the confidence level should be 100% for a true-speaker trial and 0% for an impostor trial. This can only occur when the score itself is ideal, e.g. the score is always above a threshold for true speaker trials and below the same threshold for impostor trials. In practice, the scores are never ideal, so it is impossible to obtain ideal confidence levels.

One of the simplest formulae to obtain a confidence level is using a logistic curve, but this requires the assumption the true and impostor speaker scores are distributed according to a Normal distribution. In this report I propose a new formula, called the dual Detection Error Trade-Off (DET) curve, since it represents the same information as a DET curve. This formula avoids the assumption of normally distributed target and impostor scores.

The quality of the confidences can be summarised using a metric known as the Normalised Cross Entropy(NCE). The NCE is maximised when the confidence level is equal to the a-posteriori probability of the speaker hypothesis, given the score.

Experiments were performed on the NIST 99 data. Two speaker verification systems were used to test the dual DET curve by comparing it against the logistic curve. The dual DET curve performed better than the logistic curve in adverse conditions (bad signal-to-noise ratio or short utterance length), but slightly worse in ideal conditions.

Author

Trevor C Tao
CCCID

Trevor Chi-Yuen Tao graduated from the University of Adelaide (Australia) in 2005 with a PhD in Applied Mathematics and started employment at DSTO in August 2006. His current research involves speech processing.

Contents

Glossary **xi**

1 Introduction **1**

2 Existing Confidence Estimators **1**

3 New Confidence Metric **2**

3.1 Ideal dual DET curve 2

3.2 Non-ideal dual DET curves 4

4 Relationship Between Confidence and Validation Data **4**

5 Confidence Evaluation **5**

6 Experiment **7**

7 Summary and Conclusion **9**

References **11**

Appendices

A Derivation of the Logistic Curve **16**

B Derivation of NCE **17**

Figures

1 DET curve and dual, n = 1.0 12
2 DET curve and dual, n = 1.1 12
3 DET curve and dual, n = 6.0 13
4 DET curve and dual, piecewise linear 13
5 Dual DET curve versus Logistic curve 14
6 DET curve, System A 14
7 DET curve, System A (20 speakers) 15
8 DET curve, System B 15

Tables

1 Dual DET Curve vs Logistic (System A proper NIST evaluation) 8
2 Dual DET Curve vs Logistic (System A, 20 speakers only) 8
3 Dual DET Curve vs Logistic (System B, proper NIST evaluation) 9

Glossary

DCF Detection Cost Function

DET Detection Error Trade-off

FAR False Alarm Rate

LLR Log Likelihood Ratio

MR Miss Rate

NIST National Institute of Standards and Technology

SNR Signal-to-Noise Ratio

1 Introduction

Speaker recognition is the problem of identifying people from their voices, and has important applications. For instance, it is often desirable to determine (or confirm in the case of Speaker *Verification*) the identities of various speakers in various applications, such as telephone calls. The human voice is unique, or at least very difficult to mimic successfully, so the ability to identify a speaker based on his or her voice offers better security over some other means of identification such as passwords (which can be forgotten or compromised) or physical objects (which can be stolen).

Given an audio file and a claimed speaker identity, a Speaker Verification system typically outputs a score, summarising the evidence in favour of the speaker hypothesis. One of the main difficulties with Speaker Verification systems is that the score cannot be easily interpreted - it is merely an arbitrary number where higher scores indicate stronger evidence in favour of the speaker hypothesis. In some cases, the score may have a specific meaning such as a log likelihood ratio (LLR) between target and background model, but in other situations there is no such interpretation available. For instance, if T-norm[1] is applied to an LLR, the new score can no longer be interpreted as an LLR.

The concept of confidence is being increasingly adopted in Speaker Verification systems[2, 3, 4, 5, 6]. Roughly speaking, a confidence level is intended to complement the system output, to indicate if the results are reliable. For instance, one can say e.g. the LLR is approximately 2.5 with 85% confidence, or the speaker should be rejected with 90% confidence.

The report is organised as follows: Section 2 discusses existing confidence estimators. Section 3 proposes a new estimator. Section 4 discusses the relationship between confidence estimators and the validation data set. Section 5 discusses evaluation of confidence estimators. Section 6 discusses an experiment comparing the performance of the proposed confidence estimator with that of an existing confidence estimator. Section 7 is the conclusion, summarising the report's findings.

2 Existing Confidence Estimators

One of the main difficulties with confidence estimators is that there are many interpretations of the word confidence. For instance higher confidence can reflect stronger evidence in favour of the speaker hypothesis or the "right decision" hypothesis (i.e. the system made the right decision to accept or reject the speaker[5]). Confidence can also reflect whether an LLR lies within a specified interval[2] (borrowing the idea from confidence intervals in statistics). In this paper I restrict my attention to confidence estimators that are interpretable as a probability $conf = p(H_1|E)$, where H_1 is the speaker hypothesis and E is some evidence, such as LLR score, channel type, signal-to-noise ratio (SNR) or other information. In other words, the confidence is an estimate of the probability that

the purported speaker is the actual speaker given the available evidence. In this report, I assume that the only available evidence is the LLR score.

One of the simplest confidence estimators uses Gaussian distributions to model the LLR scores from true-speaker and impostor trials[7]. Assuming that the prior probabilities π_0, π_1 of the speaker and null hypotheses are known, the confidence can be estimated using Bayes law:

$$p(H_1|E) = \frac{\pi_1 p(E|H_1)}{\pi_0 p(E|H_0) + \pi_1 p(E|H_1)} \quad . \tag{1}$$

Assuming that $p(E|H_0)$ and $p(E|H_1)$ are normally distributed with the same variance, this reduces to a logistic function

$$p(H_1|s) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 s)} \quad . \tag{2}$$

For the derivation of this see Appendix A. The parameters β_0, β_1 can be estimated via regression. Note that this bypasses the requirement of knowing the priors.

A well-known disadvantage of this approach is that the score distributions are generally not Gaussian (although this does not necessarily lead to poor results on actual data[8]). There are many other, more complex confidence estimators available. It is beyond the scope of this report to discuss these. I refer the interested reader to [5].

3 New Confidence Metric

3.1 Ideal dual DET curve

I propose a new way of measuring confidence. The key observation is that a detection cost function (DCF) of the form $cost = \alpha MR + (1-\alpha)FAR$ can be associated with a confidence level of $100(1-\alpha)\%$ in the following sense: given the above DCF, accepting the speaker hypothesis would cost more in the long run than rejecting it, unless one was at least $100(1-\alpha)\%$ confident of the speaker hypothesis. For any confidence level $100(1-\alpha)\%$ and corresponding DCF $cost = \alpha MR + (1-\alpha)FAR$ one can calculate the optimal threshold $\theta_{1-\alpha}$. This yields a function from confidence to score. Assuming this function is strictly monotonically increasing and hence invertible, one can obtain a function from score to confidence. It turns out this function is "orthogonal" to the Detection Error Trade-off (DET) curve in the sense of representing the same information in a different manner. In other words, one can obtain the confidence function from the DET curve and vice versa. Hence I call this function the *dual DET curve*. To make this mathematically precise, consider an "ideal" DET curve satisfying the following properties:

- For any false alarm rate FAR the miss rate is given by $MR = f(FAR)$ for some function $f : [0, 1] \to [0, 1]$.
- The function f is monotonically decreasing (i.e. $f(x) > f(y)$ if and only if $x < y$) with $f(0) = 1$, $f(1) = 0$.
- The function f is smooth, strictly convex (i.e. first derivative is strictly increasing) with $f'(0) = -\infty, f'(1) = 0$.

The first two properties are derived by considering the possible trade-off in a DET curve by varying the threshold: a higher false alarm rate implies a smaller miss rate (and vice versa) with the extreme values $f(0) = 1, f(1) = 0$ corresponding to a threshold of plus or minus infinity. The third property implies a unique confidence level for any false alarm rate. For any FAR it is not hard to show that the DCF $\alpha MR + (1-\alpha)FAR$ is minimised if and only if $f'(FAR) = -(1-\alpha)/\alpha$. Hence the confidence is given by

$$q(FAR) = 1 - \alpha = \frac{-f'(FAR)}{1 - f'(FAR)} \quad . \tag{3}$$

Conversely, equation (3) implies that $f'(FAR) = -q(FAR)/(1 - q(FAR))$ and thus

$$f(FAR) = 1 - \int_0^{FAR} \frac{q(t)}{1 - q(t)} dt. \tag{4}$$

Hence q can be obtained from f and vice versa.

Figures 1 to 4 demonstrate four examples of DET curves[1] and corresponding duals. The first three are a unit ball of the form $(1 - FAR)^n + (1 - MR)^n = 1$ for $n = 1.0, 1.1, 6.0$. In Figure 1, the DET curve is a perfect straight line indicating chance performance, i.e. the true and impostor speaker scores are identical. As expected there is no discriminating capability since the confidence is 50% for any FAR. Figure 2 shows a DET curve not far away from chance performance and as expected, the confidence is very poor. Even for extreme values of FAR the confidence is always between 0.3 and 0.7. In Figure 3 much better discriminability is observed. For instance, extreme values of FAR yield a confidence near 0 or 1, indicating near certainty of a decision (reject for 0, accept for 1). This is clarified in Figure 4 where the DET curve is piecewise linear instead of a unit ball. In this case, the confidence is 90% for $FAR < 0.1$ and 10% for $FAR > 0.1$.

The above discussion relates a function from FAR to MR with a function from FAR to confidence. In practice one is interested in a function from *score* to MR or confidence. Unfortunately, it is not easy to relate the score to FAR. In fact, this information is unavailable in a standard DET plot. However since the FAR is a decreasing (and hence

[1]Not all example DET curves are ideal but the results given can be obtained by approximation, e.g. treating the DET curve as a limit of a sequence of ideal DET curves.

invertible) function of score, it is reasonable to consider the confidence function as being a dual to the DET curve.

Figure 4 shows that for a good DET curve (near the bottom left corner), the confidence will be near 1 for scores above the threshold and near zero for scores below (recall that FAR is a decreasing function of score, so the effect is to flip the right diagram of Figure 4 horizontally). Thus in the ideal case, both the DET curve and dual DET curve will have their "energy" concentrated at zero, i.e. MR and confidence are both near unity when FAR is near zero, or both near zero when FAR is away from zero.

3.2 Non-ideal dual DET curves

In practice one does not work with an ideal DET curve and corresponding dual curve. The DET curve is calculated on some finite-size dataset so it is approximated by a zigzag sequence of horizontal and vertical lines. Moreover, the DET curve is not guaranteed to be convex, even if we were allowed to approximate the curve (by replacing the horizontal and vertical lines with a smooth curve). The non-convexity of the DET curve implies that there is not necessarily a 1-1 correspondence between confidence level and LLR score. In other words, given a confidence level $100(1-\alpha)\%$, there may be several thresholds that minimise the DCF $cost = \alpha MR + (1-\alpha)FAR$.

Thus one must approximate the dual DET curve as a piecewise smooth function, where specified confidence levels are chosen. If one chose, say, multiples of 5% confidence, the dual DET curve would be piecewise linear with "nodes" $(\theta_{0.05}, 0.05), (\theta_{0.10}, 0.10) \ldots (\theta_{0.95}, 0.95)$. If we wanted finer granularity we could have multiples of 1% confidence instead of 5% confidence. One must also have a rule for choosing between multiple thresholds corresponding to a particular confidence level. This is discussed in Section 6.

Note that care is taken to avoid a "hard" confidence level of 100% or 0% at all costs. In any case, the nature of statistical testing implies that one never expects a confidence of 100% or 0%. Moreover, the Normalised Cross Entropy (NCE) measure severely penalises hard confidence levels if they turn out to be wrong. This is described below in Section 5. In the case of the dual DET curve, the simplest solution is to set upper and lower bounds for confidences. Any confidence level above q_{max} or below q_{min} is adjusted to q_{max} or q_{min} for some $0 < q_{min} < q_{max} < 1$ respectively.

4 Relationship Between Confidence and Validation Data

Recall that the purpose of a confidence measure is to convert some evidence (such as an LLR score) into an estimate of how confident we are that the hypothesised speaker is the real speaker. The latter essentially depends on results obtained from some "previous"

validation data[2]. For instance, an LLR of -2.5 from a particular system may seem low, since the score was negative. Indeed, if the validation data was such that most of the targets scored positive and most impostors scored around zero, then an LLR of -2.5 indicates strong evidence in favour of the non-speaker hypothesis. But if the target and impostor scores were clustered near -2.0 and -5.0 respectively, then the LLR of -2.5 is in fact "high". Thus, roughly speaking, the confidence is a function of the actual LLR and some statistics obtained from LLR scores obtained on a previous dataset.

The essential point is that one cannot obtain the confidence from an LLR score *per se*. The conversion can only take place in the context of some "previous" validation data as alluded to above. This dataset must therefore be homogenous with respect to the LLR, in the sense that similar scores are generated. Ideally, the previous dataset will be generated by the same speaker(s) as those being tested, under similar conditions regarding noise, channel, SNR, environment, and so on. As is well known, it is not trivial to obtain a sufficient amount of homogenous data in real-world scenarios, but discussion of this is beyond the scope of this report.

5 Confidence Evaluation

It is important to have some method of confidence evaluation, i.e. a metric to compare different confidence estimators. One can think of this as a "meta-confidence", i.e. how confident we are that my confidence measure is good. A confidence metric is good if it consistently returns high confidences for true-speaker cuts and low confidences for impostor cuts.

To express this in mathematics, let $q(E)$ denote a function that estimates a confidence level given the evidence E. Let $M(q)$ denote the quality of the confidence estimator. Thus if the Normalised Cross Entropy (see below) is used as the confidence estimator, then the NCE would reflect the quality of the confidences in the same way the DET curve reflects the quality of the LLR scores.

The ideal confidence estimator would be a function $\hat{q}(E)$ that returns 1 when the speaker hypothesis is correct and 0 otherwise. This ideal estimator corresponds to perfect knowledge given the evidence, i.e. $p(H_1|E) = 1$ when the speaker hypothesis is correct, and $p(H_1|E) = 0$ when the null hypothesis is correct. The closer the actual confidence estimator q is to the ideal, the better (i.e. higher) the estimate $M(q)$.

There are many choices of function $M(q)$. I choose the well-known Normalised Cross Entropy (NCE) measure. The NCE is given by

$$NCE(q) = \frac{H(\omega) - \Delta}{H(\omega)} \quad , \tag{5}$$

[2]This point will be obvious to experienced researchers; the discussion is mainly intended for the layman.

where ω is a random variable representing the true identity (ω = accept with probability π_1, reject with probability $\pi_0 = 1 - \pi_1$). $H(\omega) = -\pi_0 \log \pi_0 - \pi_1 \log \pi_1$ is the entropy of ω. $\Delta = \mathrm{E}[H_1? - \log_2 q(E) : -\log(1 - q(E))]$ where $a?b : c$ is a shorthand for b if a is true and c otherwise[3], and $E[\cdot]$ denotes expectation with respect to $p(E, \omega)$, treating the true identity and evidence as random variables (in particular, when the only evidence taken into account is the score, then E is a random variable over the real line).

To maximise NCE, we need to minimise the term Δ, since $H(\omega)$ does not depend on q. One can show that $\Delta = \mathrm{E}[H_1? - \log_2 q(E) : -\log(1 - q(E))]$ reduces to

$$\Delta = \int_E -\pi_0 p(E|H_0) \log(1 - q(E)) - \pi_1 p(E|H_1) \log q(E) dE. \tag{6}$$

One can also show that Δ is minimised (and hence NCE maximised) when

$$q(E) = p(H_1|E). \tag{7}$$

Thus the NCE has a desirable property: the confidence estimator is encouraged to reflect the probability of the target hypothesis given the evidence. A metric with this property is called a proper scoring rule[3].

When evaluated on a dataset, the NCE is approximated by

$$\Delta = -\pi_0 \frac{1}{N_f} \sum_{i=0}^{N_f} \log(1 - q(E_i^f)) - \pi_1 \frac{1}{N_t} \sum_{i=1}^{N_t} \log q(E_i^t), \tag{8}$$

where N_f, N_t denote the number of impostor and target trials.

The proof of equations (6)-(8) is shown in Appendix B.

Equations (5) and (6) imply the NCE has an upper bound of 1, occurring only for the ideal estimator $q = \hat{q}$. In practice, the ideal estimator is generally not obtainable. For instance, if the dataset had one target and one impostor trial where the system outputs the same evidence, then it is impossible to have $q(E) = 1$ and $q(E) = 0$ for both trials, since the confidence q can only be a function of the evidence. The ideal estimator $\hat{q}$ could only occur if the evidence itself were "ideal", e.g. the LLR is always above a threshold for targets and below the same threshold for impostors.

An NCE of zero indicates "baseline" performance where the only evidence taken into account is the prior probabilities (π_0, π_1). Of course it is possible to have a poorly performing confidence estimator with NCE = 0 that uses evidence other than the prior. It is also possible for the NCE to be arbitrarily large and negative, indicating very bad performance. This occurs, for example, if the confidence estimator returns extreme values in the wrong direction (near 0 for true speakers or 1 for impostors).

[3]This notation is derived from programming languages such as C++ and Java.

6 Experiment

We compare the performance of the dual DET curve (Section 3) with that of the logistic curve (described in Section 2).

The dual DET curve is approximated as a piecewise linear function using nodal points (θ_q, q) where θ_q is the threshold corresponding to confidence level q and

$$q \in \{0.01, 0.05, 0.10, 0.15, \ldots, 0.95, 0.99\}. \tag{9}$$

Thus the nodes of the dual DET curve occur when $q = 0.01$ or 0.99 or a multiple of 0.05. I found this to be a reasonable compromise between accuracy and complexity. If there are multiple thresholds corresponding to the minimum of $\alpha MR + (1-\alpha)FAR$ I simply chose the smallest threshold. I found experimentally that this did not cause significant errors.

There is an important subtlety in estimating the parameters of the dual DET curve: the number of true and impostor scores must be assumed equal. Otherwise one would have to apply linear weighting to compensate. For instance, if there were 20,000 impostor scores but only 200 true speaker scores, one must pretend that each true speaker score occurs hundredfold instead of once. It is not necessary to enlarge the actual list of scores since the same effect can be achieved by using the following DCF:

$$cost = \frac{\alpha MR}{N_f} + \frac{(1-\alpha)FAR}{N_t} \quad , \tag{10}$$

where N_t and N_f are the number of true and impostor trials respectively. An example of the dual DET curve versus logistic function is shown in Figure 5. This shows that the dual DET curve is a reasonable estimator of confidence as a function of score, in that it closely approximates the logistic curve.

We used the NIST 98 and 99 male datasets for development and evaluation data respectively. The development data was used to estimate the parameters of the logistic curve and dual DET curve. The evaluation data was used to evaluate the quality of the confidence estimator.

We tested two systems, which are referred to as Systems A and B. System A was also tested in a more realistic scenario where 20 speakers were tested against each audio file and the *set of speakers was identical for each audio file* (this differs from the usual basic single speaker recognition task, where a different set of 11 speakers was tested for each audio and the correct speaker was usually among the 11 speakers tested). A consequence of the third experiment is that there are many more impostor trials than the former because the proper NIST specification was designed so that most cuts had the correct speaker among the eleven claimed speakers.

We tested nine conditions: (1) all trials (2) only those with long utterance length (3) only those with medium utterance length (4) only those with short utterance length (5) all* trials[4], where the length is determined and the logistic or dual DET curve corresponding to the appropriate level (GOOD OKAY BAD) is chosen. (6)-(9) is the same as (2)-(5) but using SNR instead of utterance length. For each condition, the NCE is recorded for both the dual DET and logistic curve. The definition of good, okay and bad is arbitrary. I defined them by sorting the data according to utterance length and dividing them into three sets of equal size. The NCE results of System A (proper NIST eval), Systems A (20 speakers only) and System B are given in Tables 1, 2 and 3 respectively. The corresponding DET curves are given in Figures 6, 7 and 8.

Overall the dual DET curve is slightly superior to the logistic curve. One possible explanation is that the logistic curve has more degrees of freedom. More specifically, (9) implies the dual DET curve has 21 degrees of freedom (for every value of q, the parameter θ_q represents one degree of freedom) and the logistic curve has only two.

Table 1: *Dual DET Curve vs Logistic (System A proper NIST evaluation)*

CONDITION	NCE(dual DET curve)	NCE(logistic)
ALL	0.574	0.539
GOOD length	0.697	0.705
OK length	0.609	0.596
BAD length	0.413	0.325
ALL* length	0.575	0.545
GOOD snr	0.679	0.691
OK snr	0.609	0.589
BAD snr	0.435	0.342
ALL* snr	0.575	0.542

Table 2: *Dual DET Curve vs Logistic (System A, 20 speakers only)*

CONDITION	NCE(dual DET curve)	NCE(logistic)
ALL	0.643	0.624
GOOD length	0.766	0.765
OK length	0.713	0.681
BAD length	0.428	0.377
ALL* length	0.659	0.632
GOOD snr	0.604	0.559
OK snr	0.653	0.614
BAD snr	0.634	0.613
ALL* snr	0.632	0.593

[4]The asterisk is only used to differentiate this from case (1).

Table 3: *Dual DET Curve vs Logistic (System B, proper NIST evaluation)*

CONDITION	NCE(dual DET curve)	NCE(logistic)
ALL	0.635	0.608
GOOD length	0.671	0.668
OK length	0.647	0.643
BAD length	0.579	0.516
ALL* length	0.633	0.609
GOOD snr	0.726	0.700
OK snr	0.682	0.675
BAD snr	0.503	0.471
ALL* snr	0.637	0.616

7 Summary and Conclusion

Our objective is to obtain a confidence level from an LLR score, since the latter is difficult to interpret. I interpret confidence as the probability of the speaker hypothesis being true, given the evidence. There are many ways to derive a confidence level given some evidence (LLR score, channel condition etc). One of the simplest metrics assumes that the only evidence taken into account is the score. The true and impostor score distributions are assumed Gaussian, and the confidence is a logistic function of score (this is shown using Bayes Law). This metric is used as a baseline.

We proposed a new confidence estimator. It is similar to the logistic curve in that the only evidence taken into account is the score. The main advantage is that it avoids the false assumption that true and impostor scores are distributed according to a Gaussian distribution. The proposed confidence estimator has an interesting property: it represents the same information as a standard DET curve. Thus confidence and the DET curve are inherently related. For this reason my confidence estimator is called the dual DET curve.

The dual DET curve is approximated by a piecewise linear function, with nodes corresponding to fixed confidence levels. However, the "granularity" (number of confidence levels) can be varied depending on the application. The dual DET curve performs slightly better than the logistic curve. This can be attributed to the fact the dual DET curve has more degrees of freedom. On the other hand, more data is required to estimate the parameters of the DET curve.

The confidence is a monotonically increasing function of score. This means that thresholding at a certain confidence level (e.g. accepting all trials whose confidence is at least 90%) is equivalent to thresholding at a certain level in the score domain. The difference is that the former is more meaningful, since 90% is interpretable as a probability of a concrete event (namely the speaker hypothesis being true, given the evidence) whereas an LLR admits no interpretation.

An obvious direction for future research is to study confidence measures that depend on some evidence other than score, e.g. SNR or channel type. This would raise some interesting issues. For example the monotonic relationship between confidence and score would no longer hold. If the evidence consists of, say, LLR score and SNR then one cannot find thresholds in the LLR-SNR domain corresponding to 90% confidence.

Acknowledgements

I wish to acknowledge Jeremy Waller and Darryn Smart for helpful discussion with this report.

References

1. R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification system, Digital Speech Processing 10 (1).

2. R. Vogt, S. Sridharan, M. Mason, Making confident speaker verification decisions with minimal speech, in: Interspeech 2008, Brisbane, Australia, 2008, pp. 1405–1408.

3. W. Campbell, D. Reynolds, J. Campbell, K. Brady, Estimating and evaluating confidence for forensic speaker recognition, in: ICASSP 2005, no. I, 2005, pp. 717–720.

4. J. G. M. Huggins, Confidence metrics for speaker identification, in: Proc. ICSLP, 2002, pp. 1381–1384.

5. J. Richiardi, P. Prodanov, A. Drygajlo, Speaker verification with confidence and reliability measures, Vol. 1, 2006, pp. 641–644.

6. J. Richiardi, A. Drygajlo, P. Prodanov, Confidence and reliability measures in speaker verification, Journal of the Franklin Institute 343 (6) (2006) 574–595.

7. H. Nakasone, S. Beck, Forensic automatic speaker recognition, in: Proc. ISCA workshop on speaker recognition - 2001: a speaker odyssey, 2001.

8. S. Bengio, C. Marcel, S. Marcel, J. Mariethoz, Confidence measures for multimodal identity verification, Information Fusion 3 (4) (2002) 267–276.

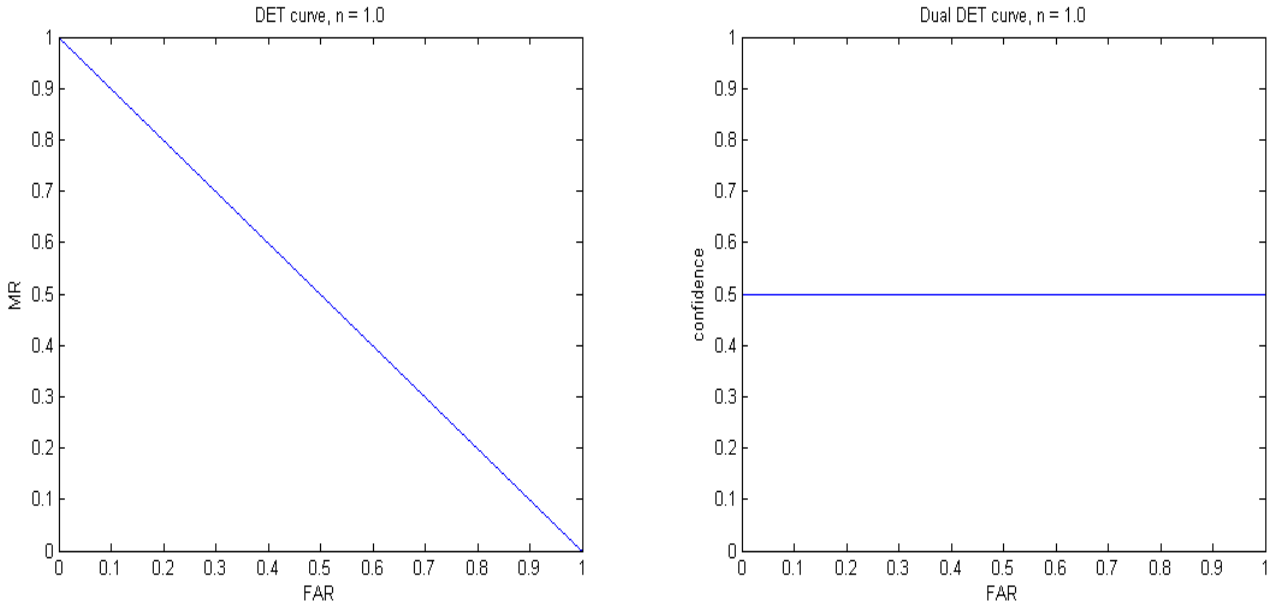

Figure 1: *DET curve and dual, $n = 1.0$*

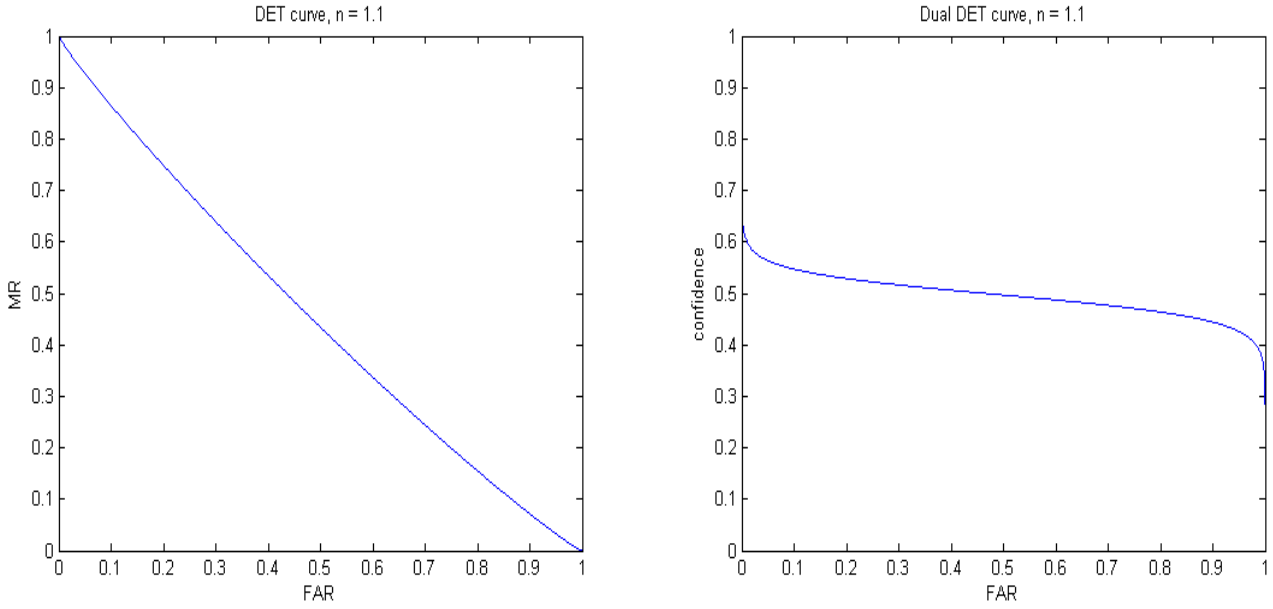

Figure 2: *DET curve and dual, $n = 1.1$*

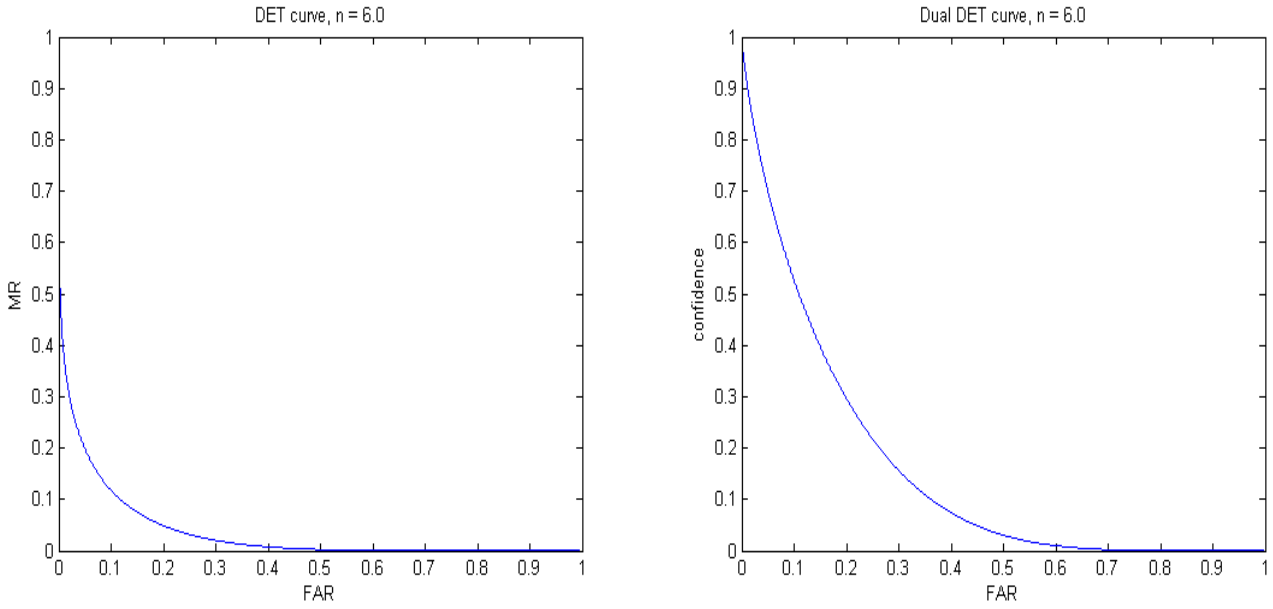

Figure 3: *DET curve and dual, $n = 6.0$*

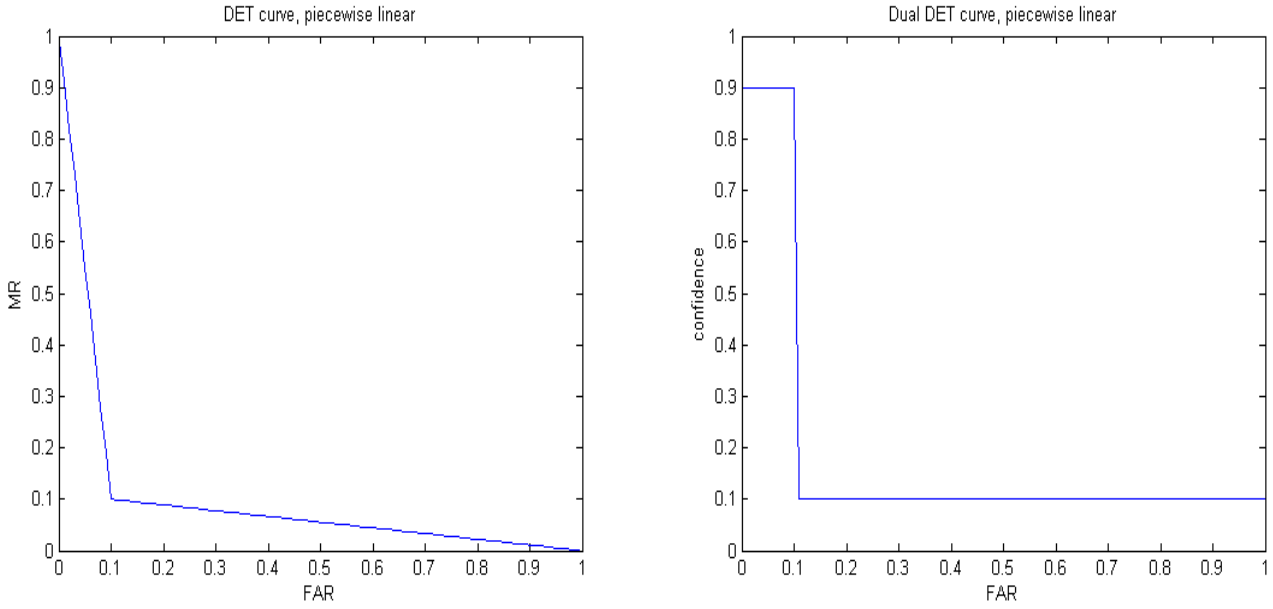

Figure 4: *DET curve and dual, piecewise linear*

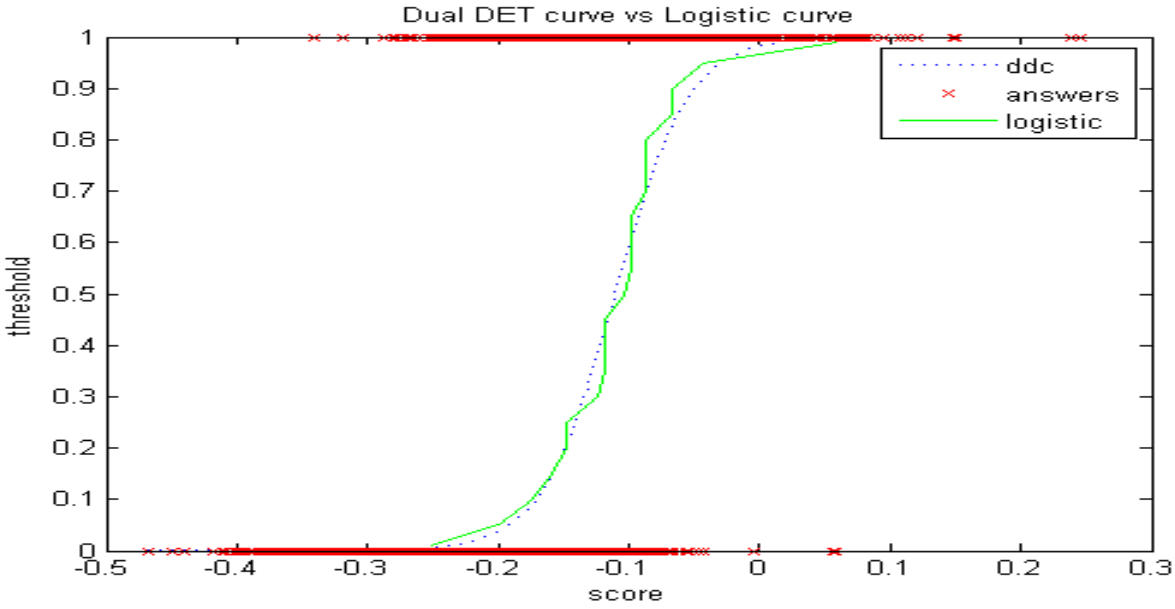

Figure 5: *Dual DET curve versus Logistic curve*

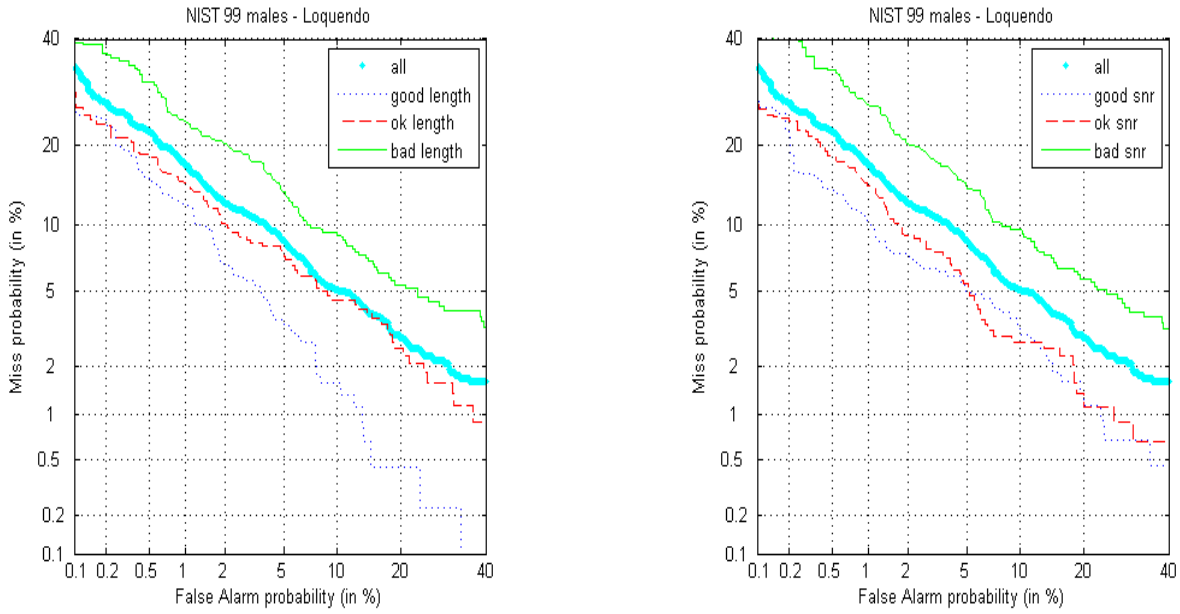

Figure 6: *DET curve, System A*

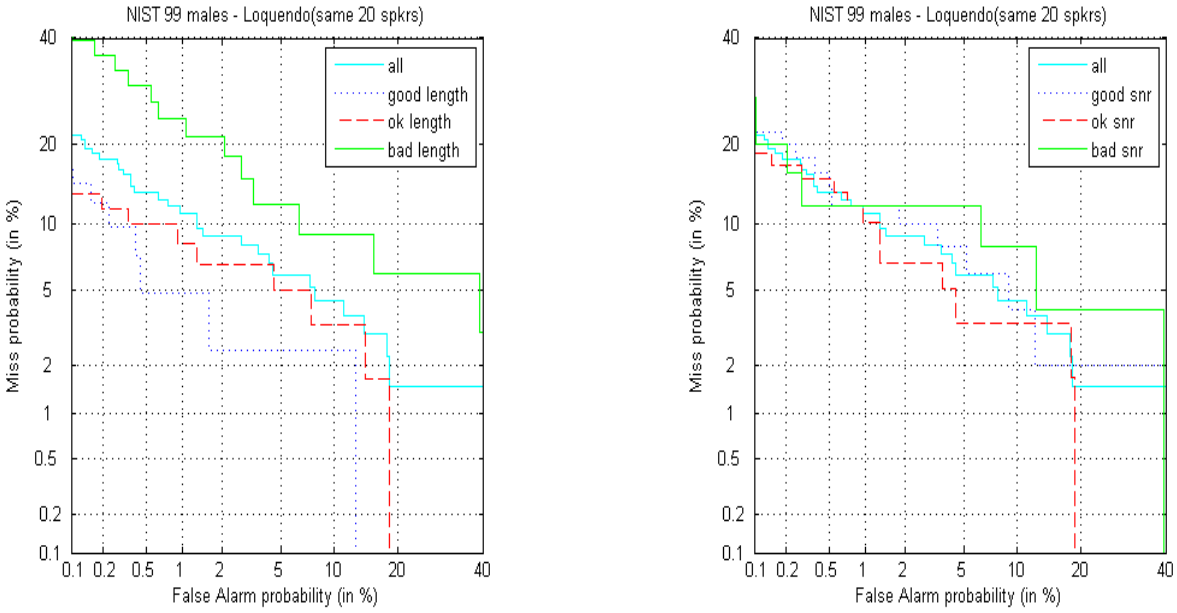

Figure 7: *DET curve, System A (20 speakers)*

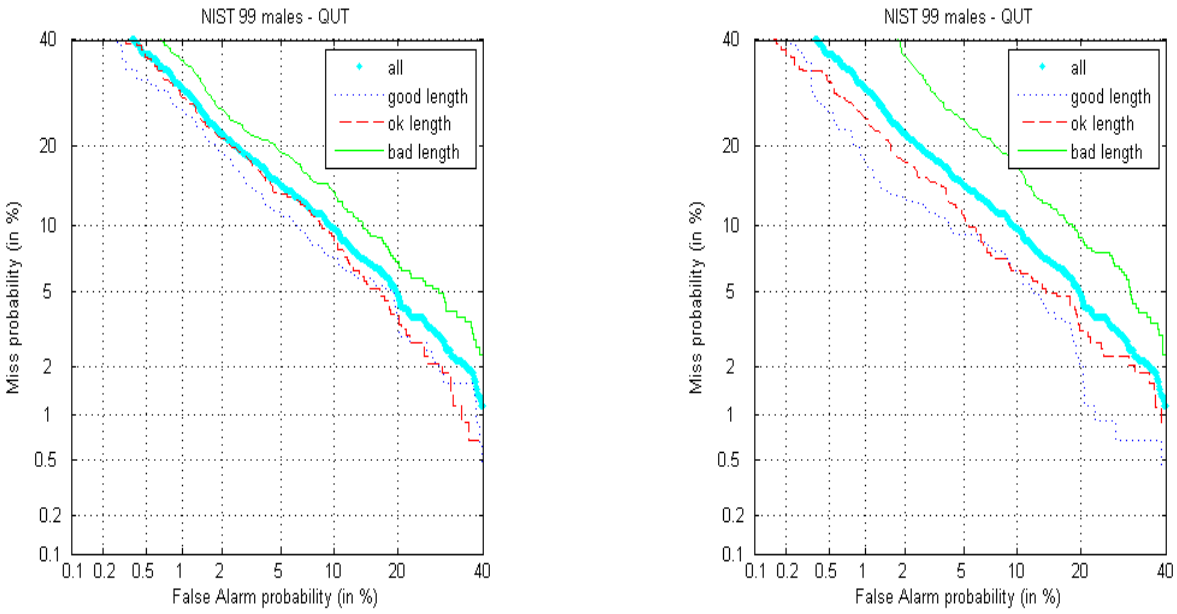

Figure 8: *DET curve, System B*

Appendix A Derivation of the Logistic Curve

Assume that

$$p(H_1|s) = \frac{\pi_1 p(s|H_1)}{\pi_0 p(s|H_0) + \pi_1 p(s|H_1)} \quad , \tag{A1}$$

where

$$p(s|H_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{(x-\mu_1)^2}{2\sigma^2} \Big), \tag{A2}$$

$$p(s|H_0) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{(x-\mu_0)^2}{2\sigma^2} \Big). \tag{A3}$$

We want to show this reduces to a logistic function

$$p(s|H_1) = \frac{1}{1+\exp(-\beta_0 - \beta_1 x)} \quad . \tag{A4}$$

Equations (A1)-(A3) imply that

$$p(H_1|s) = \frac{\pi_1 e^{-x^2/2\sigma^2} e^{2\mu_1 x/2\sigma^2} e^{-\mu_1^2/2\sigma^2}}{\pi_1 e^{-x^2/2\sigma^2} e^{2\mu_1 x/2\sigma^2} e^{-\mu_1^2/2\sigma^2} + \pi_0 e^{-x^2/2\sigma^2} e^{2\mu_0 x/2\sigma^2} e^{-\mu_0^2/2\sigma^2}} \tag{A5}$$

$$= \frac{\pi_1 e^{2\mu_1 x/2\sigma^2} e^{-\mu_1^2/2\sigma^2}}{\pi_1 e^{2\mu_1 x/2\sigma^2} e^{-\mu_1^2/2\sigma^2} + \pi_0 e^{2\mu_0 x/2\sigma^2} e^{-\mu_0^2/2\sigma^2}} \tag{A6}$$

$$= \frac{C_1 e^{D_1 x}}{C_1 e^{D_1 x} + C_0 e^{D_0 x}} \tag{A7}$$

$$= \frac{1}{1 + C_0/C_1 e^{(D_0 - D_1)x}} \tag{A8}$$

$$= \frac{1}{1+\exp(-\beta_0 - \beta_1 x)} \quad , \tag{A9}$$

where $C_1 = \pi_1 \exp(-\mu_1^2/2\sigma^2)$, $D_1 = 2\mu_1/2\sigma^2$ and similarly for C_0 and D_0, $\beta_0 = -\log(C_0/C_1)$ and $\beta_1 = (D_1 - D_0)$.

Hence $p(s|H_1)$ reduces to a logistic function, as claimed.

Appendix B Derivation of NCE

We wish to derive equations (6)-(8).

$$\begin{aligned}
\Delta &= \mathrm{E}[H_1? - \log_2 q(E) : -\log(1-q(E))] && \text{(B1)}\\
&= \int_{E,\omega} p(E,\omega)[H_1? - \log_2 q(E) : -\log(1-q(E))]d(E,\omega) && \text{(B2)}\\
&= \sum_{\omega} p(\omega) \int_E p(E|\omega)[H_1? - \log_2 q(E) : -\log(1-q(E))]dE && \text{(B3)}\\
&= -\int_E \pi_0 p(E|H_0)\log(1-q(E)) - \pi_1 p(E|H_1)\log q(E)dE, && \text{(B4)}
\end{aligned}$$

thus establishing equation (6).

To establish equation (7), one can use elementary calculus to show that the function

$$f(q) = -A\log q - B\log(1-q), \qquad 0 \le q \le 1, A > 0, B > 0 \tag{B5}$$

is minimised when $q = A/(A+B)$. Note that the convention $0\log 0 = 0$ is necessary when $A = 0$ or $B = 0$. Therefore (B4) is minimised when

$$q(E) = \frac{\pi_1 p(E|H_1)}{\pi_0 p(E|H_0) + \pi_1 p(E|H_1)} = p(H_1|E), \tag{B6}$$

using Bayes law, and hence we obtain (7).

For a given data set we approximate (6) by replacing the integration with a finite summation over a discrete set of values of E where there is at least one occurence of E. The quantity $p(E|H_1)$ can be approximated by N_{Et}/N_t where the numerator is the number of true-speaker trials where the value of E was obtained as evidence, and the denominator is the total number of true-speaker trials. The approximation $p(E|H_0) \approx N_{Ef}/N_f$ is similar.

Thus (B4) can be approximated by

$$\begin{aligned}
\Delta &\approx -\pi_0 \sum_E \frac{N_{Ef}}{N_f}\log(1-q(E)) - \pi_1 \sum_E \frac{N_{Et}}{N_t}\log q(E) && \text{(B7)}\\
&= -\pi_0 \frac{1}{N_f}\sum_{i=0}^{N_f}\log(1-q(E_i^f)) - \pi_1 \frac{1}{N_t}\sum_{i=1}^{N_t}\log q(E_i^t), && \text{(B8)}
\end{aligned}$$

and we obtain equation (8).

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA	1. CAVEAT/PRIVACY MARKING

2. TITLE	3. SECURITY CLASSIFICATION
A Confidence Estimator for Speaker Verification Using Dual DET Curves	Document (U) Title (U) Abstract (U)
4. AUTHOR	5. CORPORATE AUTHOR
T. C. Tao	Defence Science and Technology Organisation PO Box 1500 Edinburgh, South Australia 5111, Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO–RR–0358	AR-014-858	Research Report	October, 2010

8. FILE NUMBER	9. TASK NUMBER	10. SPONSOR	11. No. OF PAGES	12. No. OF REFS
2009/1137055/1	DST 97/007	CDS	17	8

13. URL OF ELECTRONIC VERSION	14. RELEASE AUTHORITY
http://www.dsto.defence.gov.au/corporate/reports/DSTO–RR–0358.pdf	Chief, Command, Control, Communications and Intelligence Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for Public Release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SOUTH AUSTRALIA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS

No Limitations

18. DSTO RESEARCH LIBRARY THESAURUS

Speech processing
Voice recognition

19. ABSTRACT

In speaker verification, the result of a trial is traditionally summarised as an arbitrary score, where a higher score indicates stronger evidence in favour of the speaker hypothesis. However this is difficult to interpret. It is useful to convert this score into a "confidence level", i.e. the posterior probability that the speaker hypothesis is correct, given the score. One of the simplest formulae to obtain a confidence level is using a logistic curve, but this requires the assumption that the true and impostor speaker scores are distributed according to a Normal distribution. In this report I propose a new formula, called the dual Detection Error Trade-Off (DET) curve, since it represents the same information as a DET curve. This formula avoids the assumption of normally distributed target and impostor scores. Experiments on the NIST 99 data prove the dual DET curve performs slightly better than the logistic curve.